# Intrusion Detection using Data Mining Techniques

Vishakha
Guru Gobind Singh Inderprasthauniversity

***Abstract:* - Intrusion detection is one of the major concerns of today's era. It is an era of computer world. Security problems in the communication world are increasing day by day. Security has become key foundation for lots of financial and web applications. Intrusion detection helps in solving network security crimes. Imperfections in IDS (intrusion detection system) gave rise to data mining in this world. In this paper we have discussed different data mining techniques for intrusion detections.**

**Keywords: data mining, intrusion detection, outlier**

## INTRODUCTION:

The internet was developed from the software called the ARPANET which was developed by U.S military. It was only restricted to military personnel and the people who had developed it. Only after it was privatized it was allowed to be used commercially. Internet is a global system of computers connected through network. For data communication certain rules are followed called as internet protocols for example TCP,UDP. There are different types of network for example LAN, MAN, WAN. Internet has changed our lives a lot. It carries lots of information.

People who belong to either private or public sector share information over internet. Information over the internet can be used by anybody. Security of internet is an important issue. In many real world applications finding intrusion detection, fraud detection and finding exceptional instances is more interesting than finding about the other crimes.

Intrusion detection system is divided in two types of approaches: misuse detection and anomaly detection.

### A. Misuse detection

In this, detection pattern are formed first and using this pattern we find out intrusion.

### B. Anomaly detection

It is defined as the expected behavior of the network traffic in advance. Any change in this pattern is defined as attack. But all changes cannot be considered as attack. The negative points of IDS are its low detection and false alarm rate.

Between these two approaches only anomaly detection algorithm has the ability to detect known attacks as misuse detections can only find out defects in known pattern.

Data mining techniques are efficient and suitable to be integrated in the intrusion detection domain, since they ensure the usage of large databases generated by the sensors.

*Data warehouse Overview:*

Bill Inmon considered being the father of Data warehousing provides the following definition:"A data warehouse is a subject oriented, integrated, non volatile and time variant collection of data in support of management's decisions."[1][2]

Dannis Murray defined "data warehouse is a collection of key pieces of information used to manage and direct the business for the most profitable outcome."[2]

Following are the description of the features of data warehouse given by Bill Inmon.

*Subject oriented data*

In operational system, we store data by individual application. For example in banking institution, data sets for a consumer loans application contained data for the particular application. Data set for other distinct application of checking account and saving account relates to those specific applications. In every industry, data set are organized around individual applications to support to those particular operation systems. These individual data set have to provide data for the specific applications to perform the specific functions efficiently. Therefore, the data sets for each application need to be organized around that specific application. In the data warehouse, data is stored by subjects not by application. Business subjects differ from enterprise to enterprise. These are subjects critical for the enterprise. For a manufacturing company, sales shipment and inventory are critical business subjects. For a retail store, sales at the checkout counter are the critical subjects

*Integrated Data*

Integrated data refers to de-duplicating information and merging it from many sources into one consistent location. When short listing our top 20 customers we must know that HAL and Hindustan aeronautics limited are same and one.

*Time Variant data*

The time variant data in data warehouse [2]

- Allows for analysis of the past
- Relates information to the present
- Enables forecasts for the future

*Non Volatile Data*

Data for the operation system are moved into the data warehouse at specific intervals. Depending on the requirements of the business, these data movements take place twice a day, once a day, once a week, or once in two weeks. Every business transaction does not update the data in the data warehouse. The business transactions update the operational system database in the real time. We add, change, or delete data from an operational system as each transactions happens but don't usually update the data in data warehouse. We don't delete the data from data warehouse in the real time. Once the data is captured in data warehouse, we don't run individual transaction to change the data there. Data updates are common place in the operational database not so in data warehouse. Data in data warehouse is not as volatile as in operational database is. The data in data warehouse is primarily for query and analysis.

### *Data Mining*

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. This information can be used to increase profit of the company revenue etc. Data mining software is one of a number of analytical tools for analyzing data. Like WEKA, MATLAB are two of them. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

KDD process is as follows [1]:

- Data preprocessing: it removes noise and inconsistent data
- Data integration: it merges data from different sources into one source
- Data selection: Useful data is taken out for analysis, un-useful data is removed.
- Data transformation: data are transformed into appropriate forms so that mining can be easier.
- Data mining: different techniques are used for getting data patterns. Different techniques are like classification, clustering etc.
- Pattern evaluation: now patterns are analyzed.
- Knowledge presentations: where visualization and knowledge representation techniques are used to present mined knowledge to users.

First four points are different forms of preprocessing. On preprocessed data, data mining is employed.

Data mining involves some common classes of tasks [3]:

- Association rule learning: – Searches for relationships between variables. It first finds frequent item set in the data set. Frequent item set are patterns that appear frequently in the data set. For example milk and tea appears together in the data set are called frequent item set.
- Clustering: – it is grouping of objects. In a group / cluster have high similarity, but they are dissimilar to objects in other groups.
- Classification: – Is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Outlier/change/deviation detection: – The identification of unusual data records, that might be interesting or data errors that require further investigation.

### *Data Mining Techniques*

#### *Clustering:*

Clustering is classification technique. It is grouping of different objects on the basis of similarities between them. Each group objects have similar property and each group is different from other by some properties. Clustering is unsupervised learning.

Clustering algorithms discover patterns and information from the data set. This discovery is based on similarity between data set objects.

K-means classifier [4]:

This classifier is based in learning from analogy, that is , by comparing a given test tuple with training test tuple that are similar to it. All test points are represented in n dimensional space. When given an unknown tuple, a K nearest neighbor classifier searches the pattern space for the k training tuples that are closest to unknown tuple. These k training tuples are the k "nearest neighbor" of unknown tuple.

#### *4.2 Decision tree [1]:*

The Decision tree classifies the given data item using the values of its attributes. The decision tree is initially constructed from a set of pre-classified data. The main approach is to select the attributes which best divides the data items into their classes. According to the values of these attributes the data items are partitioned. This process is recursively applied to each partitioned subset of the data items. The process terminates when all the data items in current subset belongs to the same class. A node of a decision tree specifies an attribute by which the data is to be partitioned. Each node has a number of edges, which are labeled according to a possible value of the attribute in the parent node. An edge connects either two nodes or a node and a leaf. Leaves are labeled with a decision value for categorization of the data.

Decision tree works even well with large amount of data. In intrusion detection there is very large amount of data set is available. Decision tree easily construct tree which is useful for intrusion detectors. Accuracy of results is another aspect. New attacks are possible due to the generalization accuracy of decision trees.

Decision tree is formed by:

- First check whether all class belongs to same class if yes than it is leaf of the tree and label that leaf.
- For every attribute calculate information gain and information, entropy.
- Find best split attribute

ID3 algorithm for decision tree: it is a supervised learning. Training set for making decision tree is used.

- Takes all attributes and calculate their entropy.

Entropy(S) = S -p (I) log2 p (I)

Where given a collection S of c outcome and p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.

Gain(S, A) is information gain of example set S on attribute A is defined as

Gain(S, A) = Entropy(S) - S ((|$S_v$| / |S|) * Entropy ($S_v$))

Where:

S is each value v of all possible values of attribute A

$S_v$ = subset of S for which attribute A has value v

|$S_v$| = number of elements in $S_v$

|S| = number of elements in S

- Choose the attribute having lowest entropy and having maximum information gain
- Make a node containing that attribute, and split over that node
- Repeat on other remaining node.

For example

Tea = {hot, cold, mild}

Sugar= {high, normal}

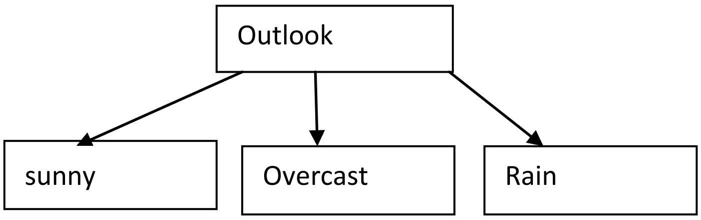
Outlook= {sunny, overcast, rain}

Let Gain(S, Outlook) = 0.24

Gain (S, Sugar) = 0.1

Gain(S, Tea) =0.024

Outlook has highest gain, so it will be used as root of the tree



Outlook used as root node so further tree depends on other two attributes that is tea and sugar.

Gain($S_{Sunny}$ ,Sugar)= 0.8

Gain($S_{overcast}$ , Temperature)=0.4

Sugar has highest gain so now it is decision node. This process repeated until all data get classified.

Genetic algorithm:

The Genetic algorithm is based on biological evolution of the life. This algorithm is inspired by mutation, inheritance and crossover.

Common algorithm [5]:

- First a initial population is created randomly.
- For each member of population fitness function is implied to find how well each population member is good for making new population.
- From that population two individual are selected whose fitness function is better than other individual, they allowed reproducing new offspring.
- Offspring thus formed mutated or crossover and then allowed to reproduce. This process keeps on repeating until got suitable solution, depends on need of programmer.

Intrusion detection for securing the network [1]

IDS:

An IDS system monitors network traffic and audit system logs in order to determine whether any violation of company standard taken place or not. It can detect intrusion that passes through the firewall or even behind the firewall.

Network attacks: the stages:

Initial uncovering:

this is 'reconnaissance' potential intruders will find out as much as they possibly can about their target by seemingly legitimate means. In this first step is finding public information from the internet about the target. Then the next is uncovering as much as possible information of company's internal network, internet domain protocol range etc. at this stage intruder cant be called as intruder because he has not violate any company security rules.

Network probe:

at this stage intruder explores more information about the company. Using 'ping sweep' of the network IP address potential targets found out, then using port scanning tool open ports found out. Even at this stage intruder cannot be called as intruder because he has not violate any company security rules.

Crossing the line toward e-crime:

Now intruder moves to commit a computer crime, by exploiting possible holes on the target machine. At this stage intruder get an administrator root privilege. This can be done through certain programming errors, by checking default login accounts by guessing easily passwords. 'Root' is basically an administrator and grants them the privilege to do anything on the system.

Capturing the network:

at this stage hackers try to own the network. Intruder will usually install some tools that replace existing files with the Trojan files. When user clicks on these files, intruder get over all control of the system thus gives a backdoor for intruders to get into control of the system.

Grab the data:

Now the intruder steals the confidential information of the user like customer card credit information. It causes potentially expensive and embarrassing situation for the organization.

Categories of intrusion detection system [6]:

Misuse detection:

Here IDS analysis the information it gathers and compares it to the database of attack signatures.

Anomaly detection:

In this system a baseline is created, it network vary from this baseline than we consider it as anomaly detection. With anomaly detection, sensors monitor network segment to compare their present state against the baseline to identify the problem.

Network based IDS:

In IDS this monitors every packets following through a point. These packets can be monitor trough firewall use. But it cannot detect attacks against a host made by an intruder who is logged in at the host's terminal.

Host based IDS:

It is installed at individual server to watch for anomalous activity. The kind of activities examined is like modification to important files, excessive CPU activity and misuse of administrative rights or root privilege.

## CONCLUSION AND FUTURE WORK:

This report based on the study of the data mining techniques for intrusion detection and how these techniques can help in intrusion detection. The objective is to help intrusion detector to understand the key characteristics of these processes and therefore select the most suitable process with respect to the type of attack.

Future work may be to consider combination of one or more techniques that may be easier than single one.

## REFERENCES

[1] paul raj Ponniah,"Data Warehouseing Fundamentals: A Comprehensive Guide for IT professionals", 2001 John Wiley & sons, Inc. ISBN:0-471-41254-6( Hardback; 0-471-22162-7(electronic)

[2] Soumendra Mohantry, Data Warehouing, Design, Development and Best Practices". Tata McGraw-Hill Publishing Company limited, new delhi

[3] Data mining," http://en.wikipedia.org/wiki/Data_mining"

[4] Surasit Songma ,Witcha Chimphlee, Kiattisak Maichalernnukul ,Parinya Sanguansat," Classification via *k*-Means Clustering and Distance-Based Outlier Detection", 2012 Tenth International Conference on ICT and Knowledge Engineering.

[5] http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

[6] Xindong Wu1 ,"Data Mining: An AI Perspective" December 2004 Vol.4 No.2, IEEE Computational Intelligence Bulletin

.